Semantic Publishing Challenge –
Assessing the Quality of Scientific Output by
Information Extraction and Interlinking⋆

Angelo Di Iorio[1], Christoph Lange[2], Anastasia Dimou[3], and Sahar Vahdati[4]

[1] Università di Bologna, Italy `diiorio@cs.unibo.it`
[2] University of Bonn & Fraunhofer IAIS, Germany `math.semantic.web@gmail.com`
[3] Ghent University – iMinds – Multimedia Lab, Belgium `anastasia.dimou@ugent.be`
[4] University of Bonn, Germany `vahdati@uni-bonn.de`

Abstract The Semantic Publishing Challenge series aims at investigating novel approaches for improving scholarly publishing using Linked Data technology. In 2014 we had bootstrapped this effort with a focus on extracting information from non-semantic publications – computer science workshop proceedings volumes and their papers – to assess their quality. The objective of this second edition was to improve information extraction but also to interlink the 2014 dataset with related ones in the LOD Cloud, thus paving the way for sophisticated end-user services.

1 Introduction: Semantic Publishing Today

The widely held assumption that 'scholarly communication by means of semantically-enhanced media-rich digital publishing is likely to have a greater impact than [print or PDF]' [1] is slowly coming true, pushed by regular events such as the workshop series on getting 'Beyond the PDF', semantic publishing and linked science[5]. Semantic technology is increasingly supporting researchers in disseminating, exploiting and evaluating their results using open formats. Concrete technical solutions investigated by the semantic publishing community include:

- machine-comprehensible representations of scientific methods, models, experiments and research data,
- links from papers to such data,
- alternative publication channels (e.g. social networks and micro-publications),
- alternative metrics for scientific quality and impact, e.g., taking into account the scientist's social network, user-generated micro-content such as discussion post, and recommendations.

Sharing scientific data and building new research on them will lead to data value chains increasingly covering the whole process of scientific research and communication. The Semantic Publishing Challenges aim at supporting the buildup of

---

⋆ The final publication is available at `link.springer.com`
[5] See `https://www.force11.org/meetings/beyond-pdf-2`, `http://sepublica.info`, and `http://linkedscience.org/category/workshop/`

such data value chains, initially by extracting information from non-semantic publications and interlinking this information with existing datasets. Our prime use case is the computation of novel quality metrics based on such information.

Section 2 presents the definition of this year's Challenge, Section 3 explains the evaluation procedure, Sections 4 to 6 explain the definitions and outcomes of the three tasks in detail, and Section 7 discusses overall lessons learnt.

2 Definition of the Challenge

In 2014, we had found it challenging to define a challenge about semantic publishing [10]. Existing datasets focused on basic bibliographical metadata or on research data specific to one scientific domain; we did not consider them suitable to enable advanced applications such as a comprehensive assessment of the quality of scientific output. We had thus designed the first Challenge to produce, by information extraction and in an objectively measurable way, an initial data collection that would be useful for future challenges and that the community can experiment on. As the two information extraction tasks had received few submissions, and as the community had asked for a more exciting task w.r.t. the future of scholarly publishing, we added an open task with a subjective evaluation.

In 2015, we left **Task 1** of 3 largely unchanged: answering queries related to the quality of workshops by computing metrics from data extracted from their proceedings, also considering information about persons and events. The 2014 results had been encouraging, and we intended to give the 2014 participants an incentive to participate once more with improved versions of their tools. As in 2014, **Task 2** focused on extracting contextual information from the full text of papers: citations, authors' affiliations, funding agencies, etc. In contrast to 2014, we now used the same data source as for Task 1 (the CEUR-WS.org open access computer science workshop proceedings), to foster synergies between the two tasks and to encourage participants to compete in both tasks. Based on the data obtained as a result of the 2014 Task 1, we defined the objective of **Task 3** to interlink the CEUR-WS.org linked data with other relevant linked datasets.

3 Common Evaluation Procedures

The evaluation for all tasks followed a common procedure similar to the other Semantic Web Evaluation Challenges:[6]

1. For each task, we initially published a *training dataset* (TD) on which the participants could test and train their extraction and interlinking tools.
2. For the information extraction tasks, we specified the basic structure of the RDF extracted from the TD source data, without prescribing a vocabulary.
3. We provided natural language queries and their expected results on TD.
4. A few days before the submission deadline, we published an *evaluation dataset* (ED), a superset of TD, which was the input for the final evaluation.

---

[6] As no one participated in Task 3, our work on this task ended with step 3.

5. We asked the participants to submit their linked data resulting from extraction or interlinking (under an open license to permit reuse), SPARQL implementations of the queries, as well as their extraction tools, as we reserved the right to inspect them.

6. We awarded prizes for the best-performing (w.r.t. the F1 score computed from precision/recall) and for the most innovative approach (determined by the chairs[7]).

7. Both before and after the submission we maintained transparency. Prospective participants were invited to ask questions, e.g. about the expected query results, which we answered publicly. After the evaluation, we made the scores and the gold standard (see below) available to the participants.

The given queries contained placeholders, e.g. 'all authors of the paper titled T'. For training, we specified the results expected after substituting certain values from TD for the variables. We evaluated by substituting further values, mostly values that were only available in ED. We defined easy as well as challenging queries, all weighted equally, to help participants get started, without sacrificing our ability to clearly distinguish the best-performing approach. A collection of PHP scripts[8] helped to automate the evaluation: they compared a CSV form of the results of the participants' SPARQL queries over their data against a gold standard of expected results, and compiled a report with measures and a list of false positives and false negatives (see Figures 1 and 2).
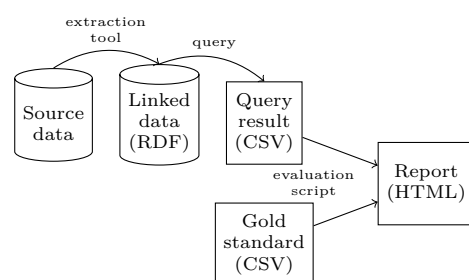

Figure 1: Precision/recall evaluation

Figure 2: Report for one query

4 Task 1: Extraction and Assessment of Workshop Proceedings Information

4.1 Motivation and Objectives

Common questions related to the quality of a scientific workshop or conference include whether a researcher should submit a paper to it or accept an invitation

---

[7] Anastasia Dimou, a co-author of one Task 1 submission [5], did not vote in this task.

[8] https://github.com/angelobo/SemPubEvaluator

to its program committee, whether a publisher should publish its proceedings, or whether a company should sponsor it [2]. Moreover, knowing the quality of an event helps to assess the quality of the papers accepted there. In the 2014 Challenge, we had designed Task 1 to extract from selected CEUR-WS.org workshop proceedings volumes RDF that would enable the computation of certain indicators for the workshops' quality [10]. The second objective of this effort was to bootstrap the publication of *all* CEUR-WS.org workshops – more than 1,400 at the time of this writing – as linked data. As discussed above in Section 2, we reused the 2014 queries, with two exceptions. As only one of the three 2014 submissions had addressed the two Task 1 queries that required metadata extraction from the PDF full text of the papers (cf. [7]), and as Task 2 focused on full-text extraction anyway, we replaced these queries (Q1.19 and Q1.20) by similar queries that only relied on information available from HTML sources.

4.2 Data Source

The input dataset for Task 1 consists of HTML documents at different levels of encoding quality and semantics.

- one HTML 4 index page linking to *all* workshop proceedings volumes (`http://ceur-ws.org/`; invalid, somewhat messy but still uniformly structured)
- the HTML tables of contents of selected volumes. Their format is largely uniform but has gained more explicit structural semantics over time, while old volumes remained unchanged. Microformat annotations were introduced with Vol-559 in 2010 and subsequently extended, to enable automatic indexing by DBLP [18]. RDFa (in addition to microformats) was introduced with Vol-994 in 2013, but its use is optional, and therefore it has been used in less than 10 % of all volumes since then. Valid HTML5 has been mandatory since Vol-1059 in 2013; before, hardly any volume was completely valid.

Challenges in processing tables of contents include the lack of standards for marking up editors' affiliations, invited talks, and further cases described in [10].

The training and evaluation datasets TD1 and ED1, available at `https://github.com/ceurws/lod/wiki/Task1`, balance different document formats. To enable reasonable quality assessment, TD1 already comprised certain complete workshop series, including, e.g., Linked Data on the Web, and, for some conferences, e.g., WWW 2012, all of its workshops that published with CEUR-WS.org. In ED1, some more workshop series and conferences were completed.

4.3 Queries

The queries were roughly ordered by increasing difficulty. Most queries from Q1.5 onward correspond to quality indicators discussed in Section 4.1; Q1.1–Q1.4 were intended to help the participants get started. Further background about Q1.1–Q1.18, which we reused from 2014, can be found in [10].

Q1.1 List the full names of all **editors of the proceedings of workshop** W.

Table 1: Task 1 Data Sources

	Training Dataset (TD1)	Evaluation Dataset (ED1)
Proceedings volumes	98	148 (98 + 50)
... including metadata of	1,700+ papers	2,400+ papers
Volumes using RDFa	6	12 (6 + 6)
... using microformats only	68	106 (68 + 38)

Q1.2 Count the **number of papers in workshop** W.

Q1.3 List the full names of all **authors who have (co-)authored a paper** in workshop W.

Q1.4 Compute the **average length of a paper** (in pages) in workshop W.

Q1.5 (publication turnaround) Find out whether the proceedings of workshop W were published on CEUR-WS.org before the workshop took place.

Q1.6 (previous editions of a workshop) Identify all editions that the workshop series titled T has published with CEUR-WS.org.

Q1.7 (chairs over the history of a workshop) Identify the full names of those chairs of the workshop series titled T that have so far been a chair in every edition of the workshop published with CEUR-WS.org.

Q1.8 (all workshops of a conference) Identify all CEUR-WS.org proceedings volumes in which workshops of conference C in year Y were published.

Q1.9 Identify those papers of workshop W that were **(co-)authored by at least one chair** of the workshop.

Q1.10 List the full names of all **authors of invited papers** in workshop W.

Q1.11 Determine the **number of editions** that the workshop series titled T has had, regardless of whether published with CEUR-WS.org.

Q1.12 (change of workshop title) Determine the title (without year) that workshop W had in its first edition.

Q1.13 (workshops that have died) Of the workshops of conference C in year Y, identify those that did not publish with CEUR-WS.org in the following year (and that therefore probably no longer took place).

Q1.14 (papers of a workshop published jointly with others) Identify the papers of the workshop titled T (which was published in a joint volume V with other workshops).

Q1.15 (editors of one workshop published jointly with others) List the full names of all editors of the proceedings of the workshop titled T (which was published in a joint volume V with other workshops).

Q1.16 Of the workshops that had editions at conference C both in year Y and $Y+1$, identify the **workshop(s) with the biggest percentage of growth** in their number of papers.

Q1.17 (change of conference affiliation) Return the acronyms of those workshops of conference C in year Y whose previous edition was co-located with a different conference series.

Q1.18 (change of workshop date) Of the workshop series titled T, identify those editions that took place more than two months later/earlier than the previous edition published with CEUR-WS.org.

Q1.19 (institutional diversity and internationality of chairs) Identify the affiliations and countries of all editors of the proceedings of workshop W.

Q1.20 (continuity of authors) Identify the full names of those authors of papers in the workshop series titled T that have so far (co-)authored a paper in every edition of the workshop published with CEUR-WS.org.

Q1.5 (partly), Q1.12, Q1.13, Q1.16 and Q1.17 relied on the main index.

As Task 1 also aimed at producing linked data that we could eventually publish at CEUR-WS.org, the participants were additionally asked to follow a uniform URI scheme: `http://ceur-ws.org/Vol-NNN/` for volumes, and `http://ceur-ws.org/Vol-NNN/#paperM` for a paper having the filename `paperM.pdf`.

4.4 Accepted Submissions and Winners

We received and accepted four submissions that met the requirements.

Milicka/Burget [11], the only new team, took advantage of the facts that, despite changes in *markup*, the visual *layout* of the proceedings volumes has hardly changed over 20 years, and that *within* one volume non-standard layout/-formatting choices are applied consistently. They do not rely on the microformat markup at all. The generic part of their data model covers a page's box layout and the segments of these boxes, which get tagged after text analysis. Further domain-specific analysis yields a logical tree structure, which is finally mapped to the desired output vocabulary. This submission won both awards: for the *most innovative approach* and for the *best performance*.

The three teams that had participated in 2014 evolved their submissions. The following overview focuses on new functionality; otherwise, we refer to the 2014 overview [10]. Kolchin et al. [8] (2014: [7]) enriched their knowledge represent-ation and optimised precision by adding post-processing steps including name disambiguation. Heyvaert et al. [5] (2014: [4]) simplified their HTML→RDF mapping definitions thanks to improvements of the RML mapping language, and optimised precision and recall by running systematic tests over the output to reduce failure due to, e.g., malformed literals. Ronzano et al. [14] (2014: [13]) consulted additional external datasets and web services to support information extraction (e.g. the EU Open Data Portal for names of institutions) and im-proved their heuristics for validating, sanitising and normalising the data extrac-ted. Their original submission performs poorly because they forgot the trailing slash of the volume URIs. We fixed this mistake to improve comparability.

4.5 Lessons Learnt

The four Task 1 submissions followed different technical approaches. Two solu-tions were solely developed to address this Challenge [8, 14], whereas Heyvaert et al. and Milicka/Burget defined task-specific mappings in an otherwise generic

Table 2: Task 1 evaluation results

Authors	Overall average precision	Overall average recall	Ov. avg. F1	Queries attempted	Average precision on these	Average recall on these	Avg. F1
Milicka/Burget [11]	0.774	0.591	0.64	1–20	0.774	0.591	0.64
Kolchin et al. [8]	0.658	0.531	0.565	1–18	0.731	0.591	0.628
Heyvaert et al. [5]	0.254	0.248	0.244	1–18, 20	0.268	0.261	0.257
Ronzano et al. [14]	0.028	0.046	0.034	1–12,	0.039	0.066	0.048
. . . with fixed URIs	0.375	0.290	0.302	14–15	0.536	0.414	0.432

Table 3: Task 1 comparison to 2014 (Q1–Q18)

2015	Average precision	Average recall	Avg. F1	2014	Average precision	Average recall	Avg. F1
Milicka/Burget [11]	0.805	0.603	0.657	n/a			
Kolchin et al. [8]	0.731	0.591	0.628	[7]	0.678	0.628	0.644
Heyvaert et al. [5]	0.283	0.276	0.271	[4]	0.153	0.103	0.117
Ronzano et al. [14]	0.031	0.051	0.037				
. . . with fixed URIs	0.417	0.322	0.336	[13]	0.372	0.348	0.319

framework [5, 11]. The performance *ranking* of the three tools evolved from 2014 has not changed (cf. Table 2), but their performance has improved (cf. Table 3) – except for Kolchin et al., who improved precision but not recall. Disregarding the two queries that were new in 2015, the tool by Kolchin et al., which had won the best performance award in 2014, performs almost as well as Milicka's/Burget's.

In 2014, we had made first experiments with rolling out the tool by Kolchin et al. at CEUR-WS.org[9], but will now also evaluate Milicka's/Burget's tool. Its reliance on the layout (which hardly ever changes) rather than the underlying markup (which improves every few years) promises low maintenance costs.

5 Task 2: Extracting contextual information from the PDF full text of the papers

5.1 Motivation and Objectives

Task 2 was designed to test the ability to extract data from the full text of the papers. It follows last year's Task 2, which focused on extracting information

---

[9] Licensing issues slowed down progress: from Vol-1265 the metadata are open under CC0, whereas for older volumes CEUR-WS.org does not have the editors' explicit permission to republish derivatives such as extracted RDF. Opinions diverge on the copyrightability of metadata [3]; DBLP actually republishes CEUR-WS.org metadata under ODC-BY. Still, CEUR-WS.org decided not to publish old metadata under their domain; instead, *we* will publish them as an outcome of this Challenge.

about citations. The rationale was that the network of citations of a paper – including papers citing it or cited by that paper – is an important dimension to assess its relevance and to contextualise it within a research area.

This year we included further *contextual information*. Scientific papers are not isolated units. Factors that directly or indirectly contribute to the origin and development of a paper include citations, the institutions the authors are affiliated to, funding agencies, and the venue where a paper was presented. Participants had to make such information explicit and exploit it to answer queries providing a deeper understanding of the context in which papers were written.

The dataset's *format* is another difference from 2014. Instead of XML sources, we used PDF this year, taken from CEUR-WS.org. PDF is still the predominant format for publishing scientific papers, despite being designed for printing. The internal structure of a PDF paper does not correspond to the logical structure of its content, rather to a sequence of layouting and formatting commands. The challenge for participants was to recover the logical structure, to extract contextual information, and to represent it as semantic assertions.

5.2 Data Source

The construction of the input datasets was driven by the idea of covering a wide spectrum of cases. The papers were selected from 21 different workshops published with CEUR-WS.org. As these workshops had defined their own rules for submissions, the dataset included papers in the LNCS and ACM formats.

Even if all papers had used the same style, their internal structures differed nevertheless. For instance, some papers used numbered citations, others used the APA or other styles. Data about authors and affiliations used heterogeneous structures, too. Furthermore, the papers used different content structures and different forms to express acknowledgements and to refer to entities in the full text (for instance, when mentioning funding, grants, projects, etc.).

The datasets TD2 (training) and ED2 (evaluation) are available at `https://github.com/ceurws/lod/wiki/Task2`, as a list of PDF files grouped by proceedings volume. Table 4 reports some statistics about these datasets. TD2 is a *randomly* chosen subset of papers from ED2. The final evaluation was performed on a randomly chosen subset of ED2 too. To cover all queries and balance results, we clustered input papers around each query and selected some of them from each cluster. Each cluster was composed of papers containing enough information to answer each query, and structuring that information in different ways.

Table 4: Task 2 Data Sources

	Training Dataset (TD2)	Evaluation Dataset (ED2)
Workshops	12	21 (12 + 9)
Papers	103 (28 ACM + 75 LNCS)	185 (103 + 22 ACM + 60 LNCS)

5.3 Queries

Our ten queries are not meant to be exhaustive but to cover a large spectrum of information. The first two collect information about authors' affiliations:

Q2.1 Identify the **affiliations** of the authors of paper X

Q2.2 Identify the **papers presented at workshop** X and written by researchers affiliated to an **organisation located in country** Y

Affiliations can be associated to authors in different ways: listed right after the author names, placed in footnotes, or placed in a dedicated space of the paper, and so on. The correct identification of affiliation and authors is tricky and opens complex issues of content normalisation and homonymity management. We adopted a simplified approach: participants were required to extract all information available in the input dataset and to normalise content.

Citations are key components of the context of a paper. Three queries deal with extracting data from bibliographies and filtering them by venue and year:

Q2.3 Identify all **works cited** by paper X

Q2.4 Identify all **works cited** by paper X and published **after year** Y.

Q2.5 Identify all **journal papers cited** by paper X

As in 2014, we some queries covered research funding. Such information is useful to investigate how funding was connected to, or even influenced, the research reported in a paper. Awareness of funding might influence the credibility and authoritativeness of a scientific work. The following two queries could be answered by parsing acknowledgements or other dedicated sections:

Q2.6 Identify the **grant(s)** that supported the research presented in paper X (or part of it)

Q2.7 Identify the **funding agencies** that funded the research presented in paper X (or part of it)

Research papers often result from large projects. Knowing them can help to better understand the scope and goal of a given work. The following query related to projects is distinct from the previous ones as these projects are peculiar and clearly identified in the papers (usually in the acknowledgements or in footnotes):

Q2.8 Identify the **EU project(s)** that supported the research presented in paper X (or part of it).

The last two queries were meant to test entity recognition from the papers' textual content. We focused on *ontologies*, as most papers in the dataset were about Semantic Web and formal reasoning and we expected ontologies to be clearly identifiable. For simplicity, we limited the search to the abstracts:

Q2.9 Identify **ontologies mentioned** in the abstract of paper X

Q2.10 Identify **ontologies introduced** in paper X (according to the abstract)

Note that we differentiated two queries: identifying all ontologies mentioned in the abstract vs. those introduced for the first time in the paper (again, search was limited to the abstract). We expected participants to analyse the text and to interpret the verbs used by the authors. Nonetheless the last five queries still proved difficult and only a few were answered correctly (see below for details).

5.4 Accepted Submissions and Winner

We received six submissions for Task 2:

Sateli/Witte [15] proposed a rule-based approach. They composed two logical components in a pipeline: a syntactic processor to identify the basic layout units and to cluster them into logical units, and a semantic processor to identify entities in text by pattern search. The framework is based on GATE and includes an RDF mapper that transforms the extracted data into RDF triples. The mapper's high flexibility contributed to this submission winning the *most innovative approach award*.

Tkaczyk/Bolikowski [19] won the *best performing tool award* for their CER-MINE framework: a Java application extracting metadata from scientific papers by supervised and unsupervised machine learning. The tool was successfully used for the Challenge with a few modifications, including the implementation of an RDF export. It performed extremely well in extracting affiliations and citations.

Klampfl/Kern [6] also used supervised and unsupervised machine learning. Their modular framework identifies and clusters building blocks of the PDF layout. Trained classifiers helped to detect the role of each block (authorship data, affiliations, etc.). The authors built an ontology of computer science concepts and exploited it for the automatic annotation of funding, grant and project data.

Ronzano et al. [14] extended their Task 1 framework to extract data from PDF. Their pipeline includes text processing and entity recognition modules and employs external services for mining PDF articles, and to increase the precision of the citation, author and affiliation extraction.

Integrating multiple techniques and services is also a key aspect of MACJa, the system presented by Nuzzolese et al. [12]. Mainly written in Python and Java, it extracts the textual content of PDF papers using PDFMiner and runs multiple analyses on that content. Named Entity Recognition (NER) techniques help to identify authors and affiliations; CrossRef APIs are queried to extract data from citations; NLP techniques, pattern matching and alignment to lexical resources are finally exploited for detecting ontologies, grants and funding agencies.

Kovriguina et al. [9] presented a simple but efficient architecture, implemented in Python and sharing code with their Task 1 submission [8]. Their approach is mainly based on templates and regular expressions and relies on some external services for improving the quality of the results (e.g., DBLP for checking authors and citations). An external module extracts the plain text from PDFs. This text is matched against a set of regular expressions to extract the relevant parts; the serialisation in RDF follows a custom ontology derived from BIBO.

Table 5 summarises the results of the performance evaluation.

Table 5: Task 2 evaluation results

Authors	Precision	Recall	F1 score
Tkaczyk/Bolikowski [19]	0.369	0.417	0.381
Klampfl/Kern [6]	0.388	0.285	0.292
Nuzzolese et al. [12]	0.274	0.251	0.257
Sateli/Witte [15]	0.3	0.252	0.247
Kovriguina et al. [9]	0.289	0.3	0.265
Ronzano et al. [14]	0.316	0.401	0.332

5.5 Lessons Learnt

We see two main reasons for the unexpectedly low performance:

The complexity of the task. When designing the task, we decided to explore a larger amount of contextual information to identify the most interesting issues in this area. In retrospect, this choice led us to defining a difficult task, which instead could have been structured differently. The queries are logically divided in two groups: queries Q2.1–Q2.5 required participants to identify logical units in PDFs; the others required additional content processing. As these two blocks required different skills, we could have separated them in two tasks. Queries within one group, however, were perceived as too heterogeneous. For next year, we are considering fewer types of queries with more cases each.

The evaluation. As we considered only some papers for the final evaluation (randomly selected among those in the evaluation dataset) some participants were penalised: their tool could have worked well on other values, which were not taken into account. Some low scores also depended on imperfections in the output format. Since the evaluation was fully automated – though the content under evaluation was normalised and minor differences were not considered errors – these imperfections impacted results negatively.

6 Task 3: Interlinking

6.1 Motivation and Objectives

Task 3 was newly designed to assess the ability to identify same entities across different datasets of the same domain, thus establishing links between these datasets. Participants had to make such links explicit and exploit them to answer comprehensive queries about events and persons. The CEUR-WS.org data in itself provide incomplete information about conferences and persons. This information can be complemented by interlinking the dataset with others to broaden the context and to allow for more reliable conclusions about the quality of scientific events and the qualification of researchers.

6.2 Data Source

The input for Task 3 consists of datasets in different RDF serialisations and different levels of encoding quality and semantics. For each dataset, we made an RDF dump and an endpoint or Triple Pattern Fragments [20] available. The complete training dataset TD3, available at `https://github.com/ceurws/lod/wiki/Task3`, comprises multiple individual datasets accessible in different ways:

CEUR-WS.org This dataset includes the workshop proceedings volumes up to Vol-1322; it was produced in January 2015 by Maxim Kolchin using his extraction tool, which had won Task 1 of the 2014 Challenge [7].

    RDF dump `https://github.com/ceurws/lod/blob/master/data/ceur-ws.ttl`
    Triple Pattern Fragments `http://data.linkeddatafragments.org/ceur-ws`

COLINDA The Conference Linked Data[10] dataset exposes metadata about scientific events (conferences and workshops) announced at EventSeer and WikiCfP[11] from 2002. COLINDA includes information about the title, description, date and venue of events. It is interlinked with DBLP (see below), Semantic Web Dog Food (see below), GeoNames and DBpedia[12].

    RDF dump `https://github.com/ceurws/lod/blob/master/data/colinda.nt`
    Endpoint `http://data.colinda.org/endpoint.html`
    Triple Pattern Fragments `http://data.linkeddatafragments.org/colinda`

DBLP The DBLP computer science bibliography [18] is the prime reference for open bibliographic information on computer science publications. It currently indexes over 2.6 million publications by more than 1.4 million authors, in more than 25,000 journal volumes, 24,000 conferences or workshops, and 17,000 monographs. We used the DBLP++ dataset[13].

    RDF dump `http://dblp.l3s.de/dblp-2015-02-14.sql.gz`
    Endpoint `http://dblp.l3s.de/d2r/sparql`
    Triple Pattern Fragments `http://data.linkeddatafragments.org/dblp`

Lancet The Semantic Lancet Triplestore dataset[14] contains metadata about papers published in the Journal of Web Semantics by Elsevier. For each paper, the dataset reports bibliographic metadata, abstract and citations.

    RDF dump `https://github.com/ceurws/lod/blob/master/data/lancet.ttl`
    Endpoint `http://data.linkeddatafragments.org/lancet`
    Triple Pattern Fragments `http://data.linkeddatafragments.org/lancet`

SWDF The Semantic Web Dog Food[15] metadata covers around 5,000 papers, 11,000 people, 3,200 organisations, 45 conferences and 230 workshops.

    RDF dump `http://data.semanticweb.org/dumps/`
    Triple Pattern Fragments `http://data.linkeddatafragments.org/dogfood`

---

[10] `http://www.colinda.org/`

[11] See `http://eventseer.net/` and `http://www.wikicfp.com/`

[12] See `http://www.geonames.org/` and `http://dbpedia.org/`

[13] `http://dblp.l3s.de/dblp++.php`

[14] `http://www.semanticlancet.eu/`

[15] `http://data.semanticweb.org/`

Springer LD This dataset[16] contains metadata of around 1,200 conference series and 8,000 proceedings volumes published by Springer in the Lecture Notes in Computer Science (LNCS), Lecture Notes in Business Information Processing (LNBIP), Communications in Computer and Information Science (CCIS), Advances in Information and Communication Technology (IFIP-AICT), and Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST) series.

RDF dump `https://github.com/ceurws/lod/blob/master/data/springer.nt`

Endpoint `http://lod.springer.com/sparql`

Triple Pattern Fragments `http://data.linkeddatafragments.org/springer`

6.3 Queries

The list of queries follows, ordered by increasing difficulty:

Q3.1 (Same entities within the CEUR-WS.org dataset) Identify and interlink same entities that appear with different URIs within the CEUR-WS.org dataset. Same persons (authors and/or editors) or same events (conferences) might have been assigned different URIs.

Q3.2 (Same entities across different datasets) Identify all different instances of the same entity in different datasets. Same entities (persons, articles, proceedings, events) might appear in different datasets with different URIs.

Q3.3 (Workshop call for papers) Link a CEUR-WS.org workshop W to its call for papers announced on EventSeer and/or WikiCfP.

Q3.4 (Workshop website) Link a workshop or conference X that appears in the CEUR-WS.org dataset to the workshop's or conference's website URL.

Q3.5 (Overall contribution to the conference) Identify all papers edited by an author A of a CEUR-WS.org paper P presented at workshop W colocated with conference C, and who was also author of a main track paper at the same conference C.

Q3.6 (Overall activity in a year) Identify, for an author A of a CEUR-WS.org paper P, all his/her activity in year Y.

Q3.7 (Full series of workshops) Identify the full series of workshop W regardless of whether individual editions published with CEUR-WS.org.

Q3.8 (Other co-authors) Identify people who co-authored with author A of a paper P published by CEUR-WS.org but did not co-author any CEUR-WS.org papers published in year Y with him/her.

6.4 Lessons Learnt

For Task 3, we did not receive any submissions, even though participants of the other two tasks had expressed interest. For the next challenge we are considering interlinking tasks that focus on directly extending the information extraction tasks. This way, we expect to lower the entrance barrier for participants of the information extraction tasks to also address interlinking.

---

[16] `http://lod.springer.com/`

7 Overall Lessons Learnt for Future Challenges

As a result of the 2014 Semantic Publishing Challenge, we had obtained an RDF dataset about the CEUR-WS.org workshops. This dataset served as the foundation to build the 2015 Challenge on. We designed all three tasks around the same dataset. This was a good choice in our opinion, as participants could extend their existing tools to perform multiple tasks, and it also opens new perspectives for future collaboration: participants' work could be extended and integrated in a shared effort for producing LOD useful for the whole community.

On the other hand, the evaluation process presented some weaknesses. One participant, for instance, suggested to use an evaluation dataset disjoint from the training dataset to avoid over-training; we should also consider a larger set of instance queries and provide users with intermediate feedback so that they can progressively refine their tools towards providing more precise results.

The definition of the tasks presented some issues this year as well. It was difficult, in particular, to define tasks that were appealing and with balanced difficulty. In retrospect, some tasks were probably too wide and difficult.

Next year, we plan to further increase the reusability of the extracted data, e.g., by asking for an explicit representation of licensing information, but primarily we want to put more emphasis on interlinking. For example, by linking publications to related social websites as SlideShare or Twitter, we will be able to more appropriately assess the impact of a scientific event within the community.

References

1. Allen, B. P. et al. Improving Future Research Communication and e-Scholarship. FORCE11 Manifesto. 2011. https://www.force11.org/about/manifesto.
2. Bryl, V. et al. What's in the proceedings? Combining publisher's and researcher's perspectives. In: *SePublica*. CEUR-WS.org 1155. 2014.
3. Coyle, K. Metadata and Copyright. In: Library Journal (2013). http://lj.libraryjournal.com/2013/02/opinion/peer-to-peer-review/metadata-and-copyright-peer-to-peer-review.
4. Dimou, A. et al. Extraction and Semantic Annotation of Workshop Proceedings in HTML using RML. In: *Semantic Web Evaluation Challenges 2014*. CCIS 457. Springer, 2014.
5. Heyvaert, P. et al. Semantically Annotating CEUR-WS Workshop Proceedings with RML. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.
6. Klampfl, S., Kern, R. Machine Learning Techniques for Automatically Extracting Contextual Information from Scientific Publications. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

7. Kolchin, M., Kozlov, F. Unstable markup: A template-based information extraction from web sites with unstable markup. In: *Semantic Web Evaluation Challenges 2014*. CCIS 457. Springer, 2014.

8. Kolchin, M. et al. CEUR-WS-LOD: Conversion of CEUR-WS Workshops to Linked Data. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

9. Kovriguina, L. et al. Metadata Extraction From Conference Proceedings Using Template-Based Approach. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

10. Lange, C., Di Iorio, A. Semantic Publishing Challenge. Assessing the Quality of Scientific Output. In: *Semantic Web Evaluation Challenges 2014*. CCIS 457. Springer, 2014.

11. Milicka, M., Burget, R. Information Extraction from Web Sources based on Multi-aspect Content Analysis. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

12. Nuzzolese, A. G., Peroni, S., Reforgiato Recupero, D. MACJa: Metadata And Citations Jailbreaker. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

13. Ronzano, F., Casamayor Del Bosque, G., Saggion, H. Semantify CEUR-WS Proceedings: towards the automatic generation of highly descriptive scholarly publishing Linked Datasets. In: *Semantic Web Evaluation Challenges 2014*. CCIS 457. Springer, 2014.

14. Ronzano, F. et al. On the automated generation of scholarly publishing Linked Datasets: the case of CEUR-WS Proceedings. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

15. Sateli, B., Witte, R. Automatic Construction of a Semantic Knowledge Base from CEUR Workshop Proceedings. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

16. Semantic Web Evaluation Challenges 2014. CCIS 457. Springer, 2014.

17. Semantic Web Evaluation Challenges 2015. CCIS. **cross-reference**. Springer, 2015.

18. The DBLP Computer Science Bibliography. `http://dblp.uni-trier.de`.

19. Tkaczyk, D., Bolikowski, L. Extracting contextual information from scientific literature using CERMINE system. In: *Semantic Web Evaluation Challenges 2015*. CCIS. **cross-reference**. Springer, 2015.

20. Verborgh, R. et al. Low-Cost Queryable Linked Data through Triple Pattern Fragments. In: *ISWC Posters & Demonstrations*. CEUR-WS.org 1272. 2014.